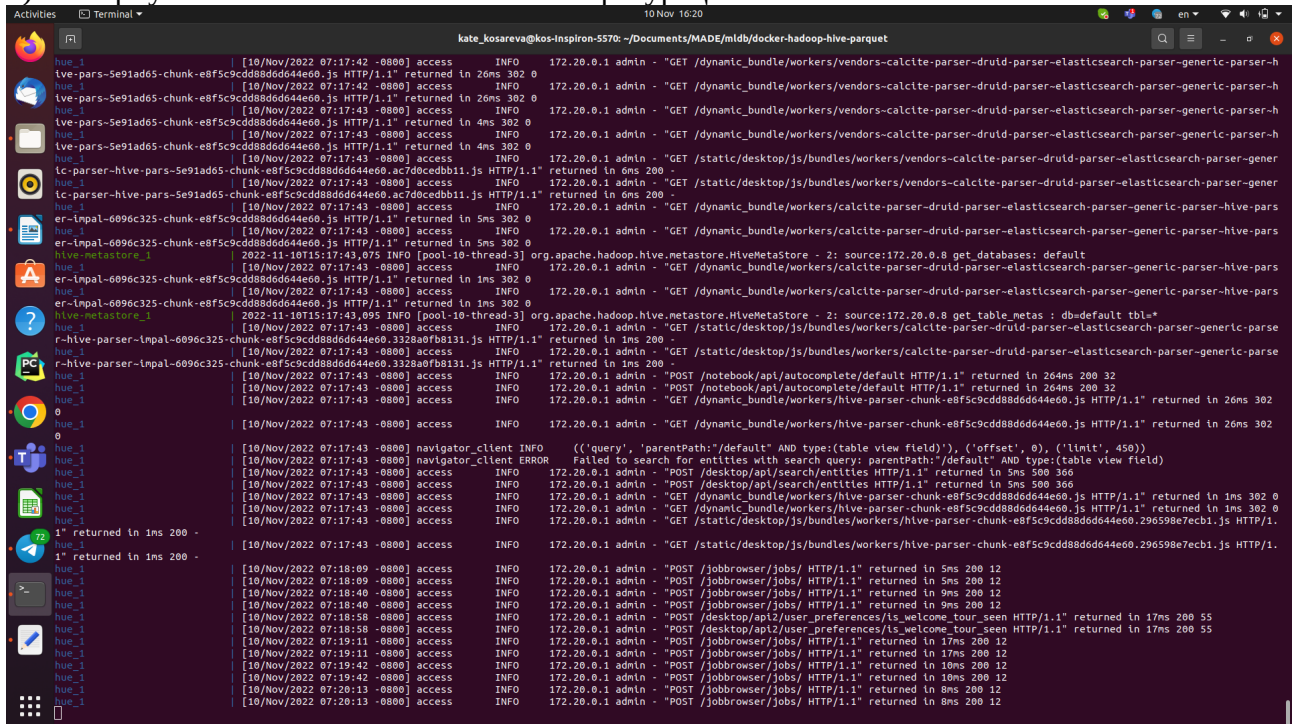


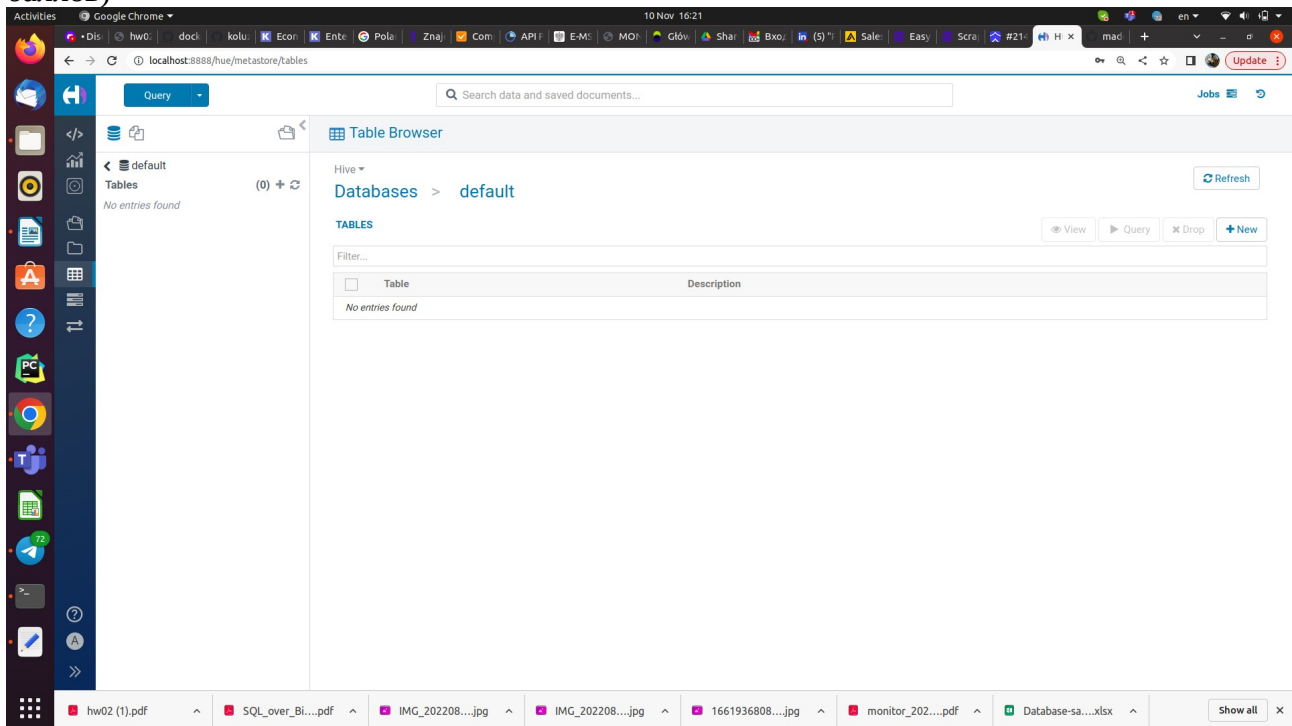
Домашнее задание 2 (Косарева Е.В, DS-12)

Блок 1

1) Развернуть локальный Hive в любой конфигурации - 20 баллов



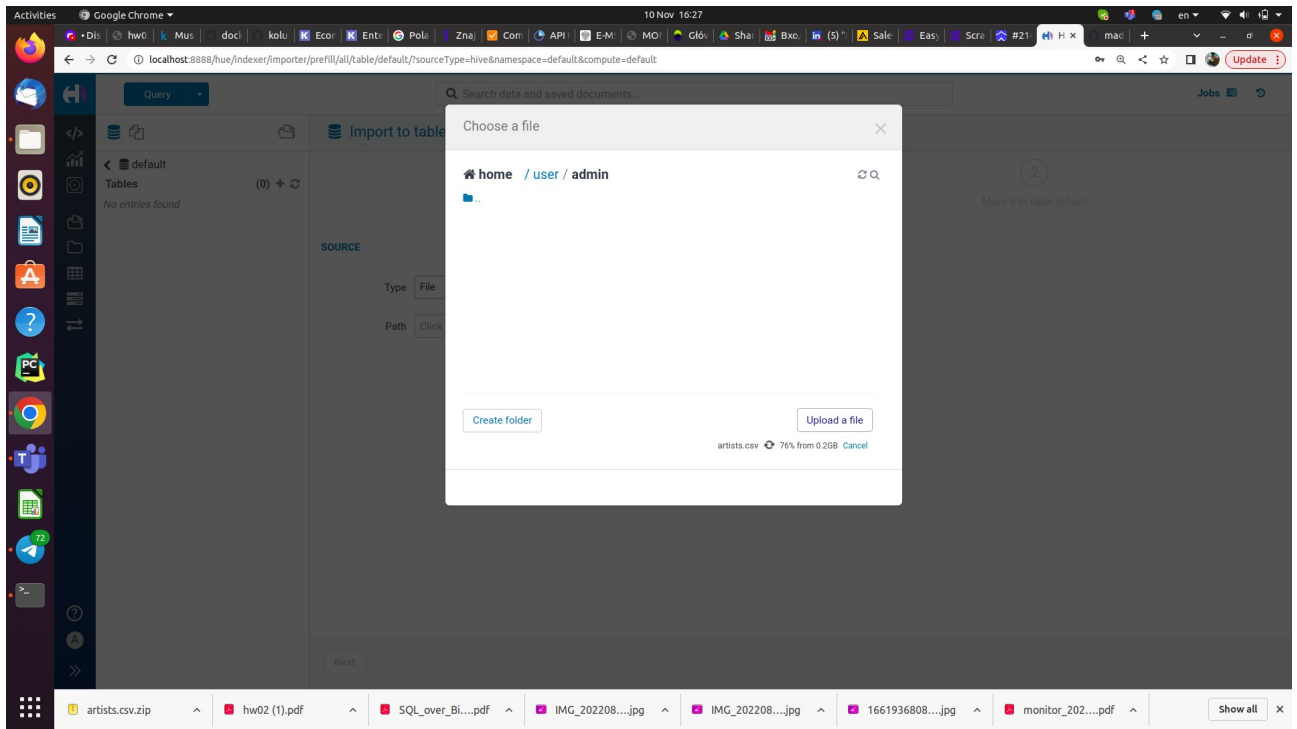
2) Подключиться к развернутому Hive с помощью любого инструмента: Hue, Python Driver, Zeppelin, любая IDE итд (15 баллов за любой инструмент, максимум 30 баллов)



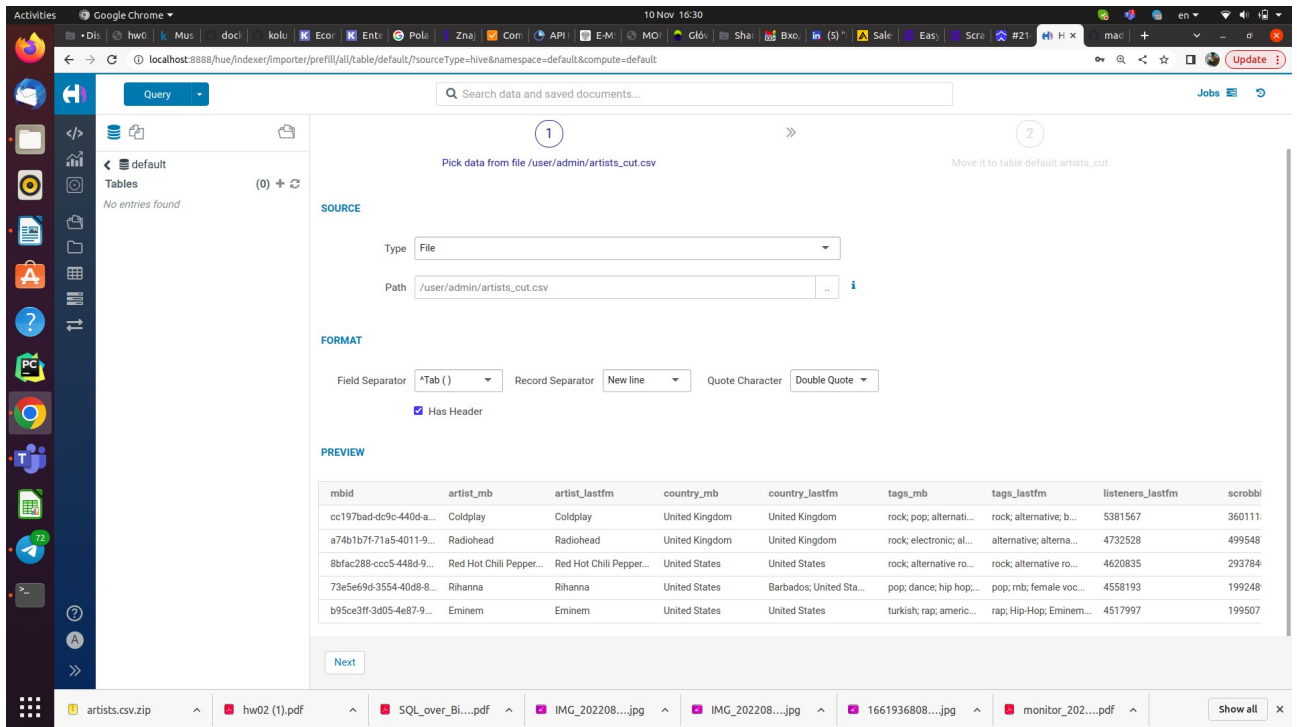
Блок 2.

1) Сделать таблицу artists в Hive и вставить туда значения, используя датасет <https://www.kaggle.com/pieca111/music-artists-popularity> - 15 баллов

Целиком файл не загружался:



Поэтому взяла часть файла



The screenshot shows the Hive Table Browser interface in a web browser. The left sidebar lists the tables in the 'default' database, including 'hue__tmp_artists_cut'. The main panel displays the 'Table Browser' for 'hue__tmp_artists_cut'. It shows the table's properties, including its location and creation time. Below this, the 'SCHEMA' section lists the columns and their data types:

Column (10)	Type	Description	Sample
mbid	string		cc197bad-dc9c-440d-a5b5-d52ba2e14234
artist_mb	string		a74b1b7f-71a5-4011-9441-d0b5e4122711
artist_lastfm	string		Coldplay
country_mb	string		United Kingdom
country_lastfm	string		United Kingdom
tags_mb	string	rock; pop; alternative rock; british; uk; britannique; britpop; pop rock...	rock; electronic; alternative rock; british; grunge; uk; britannique; brit...
tags_lastfm	string	rock; alternative; britpop; alternative rock; indie; british; seen live; Co...	alternative; alternative rock; rock; indie; electronic; seen live; british; ...
listeners_lastfm	bigint		5381567
scrobbles_lastfm	bigint		360111850
ambiguous_artist	boolean		false

2) Используя Hive найти (команды и результаты записать в файл и добавить в репозиторий):

а) Исполнителя с максимальным числом скробблов - 5 баллов

```
1 SELECT artist_lastfm FROM 'hue__tmp_artists_cut'
2 WHERE scrobbles_lastfm IN (SELECT max(scrobbles_lastfm) FROM 'hue__tmp_artists_cut');
```

The screenshot shows the Hive Query Editor interface. The query editor contains the following SQL query:

```
1 SELECT artist_lastfm FROM 'hue__tmp_artists_cut'
2 WHERE scrobbles_lastfm IN (SELECT max(scrobbles_lastfm) FROM 'hue__tmp_artists_cut');
```

The query has been executed, and the results are displayed in the 'Results (1)' tab. The result shows the artist 'The Beatles' with the highest number of scrobbles.

artist_lastfm
The Beatles

б) Самый популярный тэг на ластфм - 10 баллов

```

1 SELECT trim(tags) AS tags, count(tags) AS cnt
2 FROM hue__tmp_artists_cut
3 LATERAL VIEW explode(split(tags_lastfm, ";")) tags_names AS tags
4 GROUP BY tags ORDER BY cnt DESC LIMIT 5;
5

```

The screenshot shows the Hue web interface with a Hive query executed. The query is:

```

SELECT trim(tags) AS tags, count(tags) AS cnt
FROM hue__tmp_artists_cut
LATERAL VIEW explode(split(tags_lastfm, ";")) tags_names AS tags
GROUP BY tags ORDER BY cnt DESC LIMIT 5;

```

The results table shows the following data:

tags	cnt
1 seen live	305
2 alternative	290
3 rock	275
4 pop	231
5 00s	219

A warning message is displayed: "WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases."

с) Самые популярные исполнители 10 самых популярных тегов ластфм - 10 баллов

-

д) Любой другой инсайт на ваше усмотрение - 10 баллов

Количество записей по странам

```

1 SELECT country_mb, count(1) AS cnt FROM hue__tmp_artists_cut
2 WHERE country_mb != '' GROUP BY country_mb ORDER BY cnt DESC LIMIT 10;

```

Activities Google Chrome 10 Nov 17:45

hue-editor LanguageManual Later... mde_ml_bd/HW2 at m... hive Tutorial » Word Co... How to select particular... How to solve word count... +

localhost:8888/hue/editor/editor=32

Query

Search data and saved documents...

Jobs

Execute and watch Add a description...

2.65s Database default Type text

```
1 SELECT country_mb, count(1) AS cnt FROM hue_tmp_artists_cut
2 WHERE country_mb != '' GROUP BY country_mb ORDER BY cnt DESC LIMIT 10;
```

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.x releases.

Query History Saved Queries Query Builder Results (10)

	country_mb	cnt
1	United States	230
2	United Kingdom	99
3	Canada	21
4	Sweden	6
5	France	6
6	Australia	5
7	Germany	4
8	Ireland	3
9	Iceland	2
10	Norway	2

Tables

Filter...

- default: hue_tmp_artists_cut
 - mbid string
 - artist_mb string
 - artist_lastfm string
 - country_mb string
 - country_lastfm string
 - tags_mb string
 - tags_lastfm string
 - listeners_lastfm bigint
 - scrobbles_lastfm bigint
 - ambiguous_artist boolean